SEKYRA FOUNDATION
Ethics & Bioethics (in Central Europe), 2025, 15 (3–4), 246–258 DOI:10.2478/ebce-2025-0015

Sentencing algorithms and equal consideration of interests

Tomislav Bracanović[1]

Abstract
This paper examines whether sentencing algorithms – machine-learning-based tools for assessing the likelihood that a convicted individual will commit further offenses if released on parole – are consistent with Peter Singer's preference utilitarianism and the principle of equal consideration of interests. It begins by explaining the functioning and ethical challenges of such algorithms, especially the challenge of individualized sentencing. The paper then explores how these algorithms align with Singer's preference utilitarianism, particularly his principle of equal consideration of interests. Analyzing the key elements of this principle – maximizing and equally weighing interests, impartiality, and rejection of irrelevant group memberships – reveals how critics might use it to oppose the implementation of sentencing algorithms. A contextually more sensitive reading of Singer's views suggests that the same principle, in fact, supports the use of these algorithms. The paper concludes that sentencing algorithms are not only consistent with Singer's position but are, in many respects, reinforced by it.

Keywords: sentencing algorithms, individualized sentencing, preference utilitarianism, equal consideration of interests

Introduction

The purpose of this paper is to analyze whether sentencing algorithms – machine-learning-based tools for assessing the likelihood that a convicted individual will commit further offenses if released on parole – are consistent with Peter Singer's preference utilitarianism and the principle of equal consideration of interests. Such an analysis can be justified by at least three reasons. The first reason is that sentencing algorithms have become a prominent subject of ethical discussion. The advancement of AI technology appears to be leading toward a situation in which specific tasks of human judges are increasingly supported (or even substituted) by algorithmic systems and in which it becomes difficult to identify strong technical or moral grounds for resisting such a trend. The second reason is that in existing research on sentencing algorithms, Singer's specific position is not typically taken into account. It is thus worth examining to what extent his ethical framework may serve as a basis for either a positive or a negative evaluation of their use, especially given that Singer's ethical views have proven to be a fruitful and influential standard for addressing a wide variety of moral challenges, from animal welfare and medical ethics to global poverty alleviation and climate change.[2] The third reason – which, in a way, connects the first two – can be found in Singer's (2011, p. ix) distinctive claim that the belief in human superiority is so deeply ingrained in our thinking across a broad range of sensitive domains that challenging it is "no trivial matter" and bound to "provoke a strong reaction". "Human superiority" attitude undoubtedly also underlies much of the resistance to AI tools like sentencing algorithms, reflected in the conviction that human judgment remains irreplaceable at the end of the day. Drawing on a network of Singer's arguments, this paper aims to demonstrate that such an attitude may be unfounded.

The paper first introduces sentencing algorithms and the problem of individualized sentencing, then examines this problem through the lens of Singer's preference utilitarianism

[1] University of Zagreb (Croatia); email: tbracanovic@ifzg.hr; ORCID: 0000-0001-8168-2194

[2] Since Singer has not systematically addressed the ethics of artificial intelligence, this article can also be read as a modest contribution to Singer scholarship in that domain. Singer's writings on artificial intelligence are either relatively short pieces (some of which are reprinted in Singer, 2023) or primarily focused on the moral status of animals and animal protection (cf. Ghose et al., 2024; Singer & Tse, 2023; Hagendorff et al., 2023). It is interesting to note, however, that Singer has his own AI version: an online persona created through dialogue with Singer, designed to replicate his philosophical views (cf. Ghose, Häyry & Singer, 2025).

and the principle of equal consideration of interests, and finally argues that, when considered in light of Singer's broader ethics, the use of such algorithms can be ethically justified.

Sentencing algorithms and the individualized sentencing problem

"Sentencing algorithms" is a shorthand for tools increasingly used in courts to assess the likelihood that an individual, already convicted of a particular crime, will commit further offenses if released on parole. Such tools are typically based on machine learning technology, a subfield of artificial intelligence that designs systems capable of autonomously detecting patterns or regularities in the data they are trained on and then applying what they have learned to new data and cases. Similar tools are extensively used across various fields – for predicting market trends, disease risks, fraudulent transactions, subscription cancellations, energy consumption, or employees at risk of leaving (cf. Marr & Ward, 2019) – and it would be surprising to see them not being introduced to courts, which inherently deal with assessing risks and making high-stakes decisions.[3]

Sentencing algorithms classify offenders using a range of data. Some of these data are static, such as age, criminal history, prior arrests, or history of violent offenses. Others are dynamic, such as substance abuse, employment status, gang affiliation, childhood abuse, or criminal attitudes. Combining all the data creates a unique profile for each offender, which is then assessed using criteria or patterns identified through machine learning analysis of a large number of past offenders. If an offender's profile closely matches those of individuals who went on to commit further crimes (after release on parole), their risk score is high, making a denial of parole more likely. If their profile is more similar to that of offenders who did not reoffend under comparable circumstances, their risk score is low, which increases the likelihood of a parole being granted. The expectation is that sentencing algorithms introduced in courts will achieve a high level of calibration – meaning that their risk predictions will reliably correspond to actual outcomes – and thus generate a number of benefits, primarily in processing large volumes of data and individual cases more quickly and accurately than judges or parole boards. However, despite this optimism, sentencing algorithms, still in the early stages of adoption, also raise several concerns and criticisms.

A frequently raised concern about sentencing algorithms is the lack of transparency. Although transparency is a common problem in most machine learning-based predictive tools, it appears particularly acute in the judicial context. Recall that sentencing algorithms assess each offender against patterns identified through machine learning analysis of a large number of past offenders. The patterns uncovered by machine learning – due to the very nature of this technology – are sometimes difficult to discern, as they involve processing large volumes of data in ways that are hard to reconstruct from a human perspective. This lack of transparency makes it challenging, if not impossible, for judges, defendants, or legal counsel to understand the reasons behind a given assessment. Transparency, however, is crucial in judicial contexts, as all parties involved must be able to understand how a particular decision was reached, especially if they wish to challenge it.

A closely related criticism of sentencing algorithms is that they are biased, meaning they produce risk assessments that unjustifiably categorize some groups – especially those defined by race or ethnicity – as more risky. "Bias" should not be taken here as suggesting that the algorithm itself holds racial or ethnic prejudice, nor that its designers intentionally created it to deliver discriminatory predictions. The "bias" refers to the suspicion that the sentencing

[3] Algorithms already in use are COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), CORELS (Certifiable Optimal Rule Lists), PSA (Public Safety Assessment), HART (Harm Assessment Risk Tool), LS/CMI (Level of Service/Case Management Inventory) and OxRec (Oxford Risk of Recidivism Tool). While some of these algorithms have been the subject of extensive discussion in academic literature and the media, it is important to emphasize that this paper does not specifically address any one of them.

algorithms may have been trained on data that is either unrepresentative or generated in problematic ways. For example, if the algorithm uses data from convicted prisoners during a period when certain racial or ethnic prejudices were unjustly widespread in courts, it would most likely perpetuate that same injustice. In more technical terms, even well-calibrated sentencing algorithms (algorithms whose risk scores accurately correspond to actual outcomes across the total population of parole applicants) may still yield incorrect risk predictions across different subgroups (sometimes referred to as "disparate error rates" or "subgroup error rates"). These are situations in which certain demographic subgroups of parole applicants may be systematically misclassified, with some being unjustifiably and disproportionately labeled as high-risk (too many false positives) and others as low-risk (too many false negatives), despite the algorithm's overall accuracy.[4]

In addition to the concerns and objections outlined, an objection often raised against sentencing algorithms (also the focus of this paper) is commonly referred to as the individualized sentencing objection. It is a problem that will most likely persist even if issues of transparency and bias are resolved (for example, if we develop algorithms using unproblematic data or with a reliable "explain yourself" button) due to the simple fact that the decision-making is done by algorithms relying on statistics, probabilities, and patterns within groups. Such algorithms, one could argue, would still violate the fundamental ethical and legal principle according to which judicial decisions must be based on assessing the offender's unique individual characteristics rather than his level of similarity to specially devised groups of past offenders. Tools like sentencing algorithms, according to Siegel, appear to "fly in the face of the very notion of judging a person as an individual" because it seems "unfair to predict a person's risk of bad behavior based on what other people – who share certain characteristics with that person – have done" (Siegel, 2018, p. 80). In his formulation of this problem, Chiao acknowledges that there is a "powerful intuition" that "showing that something is true of many people just like you is very different from showing that it is true of you" (Chiao, 2023, p. 21).[5] In 2014, U.S. Attorney General Eric Holder expressed concerns that although introduced with good intentions, such tools might "inadvertently undermine our efforts to ensure individualized and equal justice" (Angwin et al., 2016). In other words, even the best-designed sentencing algorithms may be fundamentally unjust due to their inconsistency with the requirement of individualized sentencing. In what follows, they will be assessed within Singer's preference utilitarianism and its principle of equal consideration of interests.

An equal consideration-based algorithms-skepticism

The core of Singer's preference utilitarianism is neatly captured in his formulation of the principle of equal consideration of interests:

> The essence of the principle of equal consideration of interests is that we give equal weight in our moral deliberations to the like interests of all those affected by our actions. […] The principle of equal consideration of interests acts like a pair of scales, weighing interests impartially. True scales favour the side where the interest is stronger or where

[4] Bias and transparency are standard textbook concerns regarding machine learning-based predictive tools in various areas of life (cf. Zerilli et al., 2021, ch. 2 and 3). A well-known bias-oriented criticism of sentencing algorithms is Angwin et al. 2016 (for a detailed response cf. Flores et al., 2017). It is not unreasonable to expect such concerns to be mitigated through the development of future, more advanced and better-audited algorithms (cf. Reich & Vijaykumar, 2021).

[5] The general argument of the present paper partially builds on Chiao's (2023) defense of algorithmic tools – not so much on its epistemological strategy, but rather on its normative strategy, which counters the individualized decision-making demand with two, legally, equally important demands: that decision-making needs to reach accurate outcomes and that law is applied consistently.

> several interests combine to outweigh a smaller number of similar interests, but they take no account of whose interests they are weighing (Singer, 2011, pp. 20–21).

Although sentencing algorithms, on the one hand, in virtue of being designed to promote interests such as public safety and judicial efficiency, obviously reflect the basic utilitarian requirement of promoting overall well-being, they also, on the other hand, seem seriously constrained by principles such as equality before the law and the right to due process, aligning with the specific demands of equal consideration of individual interests. In order to demonstrate the potentially constraining nature of the principle of equal consideration of interests, Singer's formulation will be divided into four requirements – maximization and equal weight of interests, impartiality, and the rejection of irrelevant groups – to demonstrate how each might pose a threat to the ethical viability of sentencing algorithms. The analysis will be presented through a hypothetical scenario: An individual serving a sentence for armed robbery – call him Mr. Blue – applies for parole but is denied based on the recommendation of a sentencing algorithm. His legal team is now preparing an appeal. In addition to their official legal actions, they drafted the following public statement structured around the four requirements of Singer's principle of equal consideration of interests.

Maximization of interests. The requirement of interest maximization, according to which more substantial or numerous interests should outweigh weaker or fewer ones, is a *prima facie* acceptable moral guideline. However, the question arises whether this requirement has been appropriately respected in the present case. The determination that Mr. Blue is a high-risk individual was made in such a way, namely, that it remains unclear which specific interests – or a combination of specific interests – were deemed to outweigh his two significant interests: his interest in being released early from prison and his interest in an individualized, rather than algorithmic or merely statistical, assessment of risk. To be sure, one can acknowledge the potential harm that any offender released on parole might cause. *Potential* harm, however, is not equivalent to *actual* harm.[6] It is, by definition, a probabilistic prediction that may or may not materialize. The harm, however, that will undoubtedly occur is the one Mr. Blue will suffer: first, by being forced to serve additional years in prison, and second, by the fact that the decision affecting his liberty is, *de facto* if not *de jure*, made by a non-human algorithm rather than by a judge or parole board. This may represent an additional layer of psychological and moral harm to Mr. Blue and his family and community (cf. Davies & Douglas, 2022, p. 105). He was not treated as a unique and morally significant individual, and this failure, viewed through a utilitarian lens, adds to the total amount of suffering he experiences without yielding any identifiable or proportionate benefit on the other side.

Equal weight. Building on the previous point, a further concern can be raised. The requirement embedded in Singer's principle that similar interests should be given equal weight is undoubtedly both morally and legally valid. However, it is debatable whether all the interests at stake in Mr. Blue's case were truly comparable – or even commensurable. Mr. Blue's interest in being granted parole and receiving an individualized assessment was not given equal weight relative to other interests in this decision-making process. The parole board's reliance on a sentencing algorithm – a tool primarily designed to increase judicial efficiency and reduce costs – suggests that specific institutional interests were prioritized over individual ones. These tools serve not only offenders seeking parole but also the courts' interests in facilitating their

[6] At this point, Mr. Blue's legal team could invoke Singer's (2011, p. 154) distinctions from his discussion on the permissibility of abortion, where he differentiates between various potential and actual states of affairs: "There is no rule that says that a potential X has the same value as an X or has all the rights of an X. There are many examples to show just the contrary. To pull out a sprouting acorn is not the same as cutting down a venerable oak. To drop a fertile egg into a pot of boiling water is very different from doing the same to a live chicken. Prince Charles is (at the time of writing) a potential king of England, but he does not now have the rights of a king".

operations and the state's interest in judicial system efficiency. Mr. Blue had a vested interest in a fully individualized evaluation, as he believed such a process would result in a more favorable outcome for him. However, this interest was subordinated to institutional cost-effective concerns. Courts, of course, serve the community, but they also, in a way, have to serve individuals brought before them by evaluating them as distinct moral persons. No matter how burdensome, slow, or complicated judicial proceedings may be, there is no moral justification for placing the interests of procedural expediency above the interest of individuals in being judged fairly based on their unique characteristics and circumstances. If this proves to be unfeasible within the current institutional framework, then it may well be that this institutional framework has to be reformed.

Impartiality. It is nearly redundant to say that interests should be weighed impartially and that the same standard must apply to sentencing algorithms. However, how confident can we be that the algorithm used in Mr. Blue's case was impartial? Recall the concerns associated with sentencing algorithms, especially the risk of them being biased. Such tools have often been criticized for relying on flawed or unrepresentative training data, and we cannot be certain that sufficient progress has been made to render them genuinely unbiased and, by extension, impartial.[7] Again, this is not to suggest that the algorithm in question was deliberately designed to be biased. Rather, the concern lies in the nature of the technology itself: these are highly complex, often opaque systems that detect and utilize patterns across a multitude of variables – many of which are undetectable to human evaluators. These patterns may encode latent biases that negatively affect individuals like Mr. Blue. Even as these tools evolve and explicitly exclude inadmissible features such as race or ethnicity, the possibility remains that other, more subtle forms of bias persist in ways we do not yet understand. If the principle of equal consideration of interests "acts like a pair of scales, weighing interests impartially" and "takes no account of whose interests they are weighing" (Singer, 2011, pp. 20–21), then sentencing algorithms may be the very antithesis of that principle.

No irrelevant groups. One can agree that the moral weight of interests should not depend on characteristics morally irrelevant to those interests, such as race, sex, or IQ. And yet, from this perspective, using a sentencing algorithm in the case of Mr. Blue – as in any comparable case – raises a difficulty. Although Mr. Blue's assessment was not based on his membership in any historically marginalized (and legally inadmissible) group, the algorithm evaluated him through comparisons with various statistical groups, each arguably irrelevant to his specific circumstances. Imagine that Mr. Blue, when compared with individuals who reoffended while on parole, exhibits a 60% similarity in specific criminal history, a 70% similarity in prior arrests, and an 80% similarity in criminal attitudes. These percentages, however, are merely statistical aggregates: they reflect patterns across groups but tell us little – if anything – about Mr. Blue as a distinct individual. One could just as reasonably point out that he *differs* from reoffenders by 40% in criminal history, 30% in prior arrests, and 20% in criminal attitudes. Who can guarantee that Mr. Blue does not belong to the more favorable subset for each of these statistical clusters? One could plausibly argue that the statistical basis of sentencing algorithms may conceal just as much as it reveals – precisely the kind of epistemic uncertainty that renders individualized assessment morally and legally indispensable. To illustrate this concern, imagine that Mr. Blue is applying to become a military pilot, and the military must assess his weight. However, instead of weighing him directly, the military predicts his weight by comparing his various other traits (typically correlated with body weight, such as height, body fat percentage, muscle mass, or metabolic rate) to those of statistically similar individuals. Most would

[7] Angwin et al. (2016) argued that certain risk assessment tools exhibited hidden negative bias against African Americans while displaying positive bias toward white Americans. Similarly, in the famous case of *State v. Loomis*, the defense argued, among other things, that the algorithm in question demonstrated negative bias against men and favorable bias toward women.

consider such a procedure unreliable and fundamentally unfair, as it disregards Mr. Blue's individuality in favor of a group-based approximation. Similarly, evaluating his risk of reoffending through algorithmic comparisons – rather than examining his particular circumstances – fails to meet the moral and legal requirement of individualized treatment.

An equal consideration-based support for algorithms

Decisions regarding recidivism and parole carry serious weight. The difference between immediate release and spending another five or ten years in prison is both existentially and morally significant for those affected, which is why using sentencing algorithms raises concerns. At first glance, the earlier discussion may suggest that such algorithms are in strong tension with the principle of equal consideration of interests. In this section, it will be argued that this is not the case. By reexamining four key requirements of Singer's principle within the broader context of his work, it will be shown that sentencing algorithms are compatible with, or even supported by, his preference utilitarianism.

Maximization of interests. Utilitarian justification for laws and legal institutions is typically grounded in their beneficial consequences. Singer is faithful to this tradition when he writes that "human beings are social in nature, but not so social that we do not need to protect ourselves against the risk of being assaulted or killed by our fellow humans" (Singer, 2011, pp. 262–263). He argues that any settled decision-making procedure for resolving disputes "economically and speedily" is preferable to using force and the harm it inevitably brings. While Singer allows for breaking the law (civil disobedience) under specific circumstances, he generally views the law as a valuable institution. He considers compliance with it to be morally warranted on two grounds. First, it fosters respect for the law and its social utility; second, law violations impose additional costs on the judicial and law enforcement system (Singer, 2011, p. 263). Sentencing algorithms can be easily accommodated into this utilitarian picture as long as they help the system, when faced with cases like that of Mr. Blue, achieve its goals "economically and speedily". The first question here is whether the benefits of such algorithms are probable enough to warrant their implementation. As Singer himself says

> Any consequentialist ethics must take probability of outcome into account. A course of action that will certainly produce some benefit is to be preferred to an alternative course that may lead to a slightly larger benefit but is equally likely to result in no benefit at all. Only if the greater magnitude of the uncertain benefit outweighs its uncertainty should we choose it. Better one certain unit of benefit than a 10 percent chance of five units; but better a 50 percent chance of three units than a single certain unit. The same principle applies when we are trying to avoid evils (Singer, 2011, p. 207).

Here is why sentencing algorithms seem to fit well within this formula. Although the full extent of their benefits may become evident only after their broader implementation, the prevailing view seems to be that these tools will be either significantly more accurate than judges and parole boards or, at the very least, as accurate as they are. From the utilitarian perspective, even such divided projections are sufficient to justify their implementation. In the worst-case scenario, the algorithms will provide predictions as reliable and precise as those made by judges and parole boards, thus freeing up their time and resources for other, more complex tasks. In the best-case scenario, their predictions will be more accurate than those made by judges and parole boards while still generating the same practical benefits regarding efficiency and resource allocation.[8]

[8] Schwarze and Roberts (2022, p. 212) warn that sentencing algorithms may fail to capture critical qualitative features, such as distinguishing between an offender who genuinely feels remorse and one who merely expresses remorse through a lawyer. However, this does not apply to all judicial decisions, especially recidivism assesment.

The utilitarian rationale for sentencing algorithms can be illustrated by an example from legal history mentioned by Schauer (2003). In the late 1980s, the *Federal Sentencing Guidelines* were introduced in the United States to reduce judicial discretion (the authority of judges to make decisions based on personal judgment) and address sentencing disparities (different sentences for similar offenses imposed by different judges). A "Sentencing Table" was implemented, combining 43 offense levels with six criminal history categories, producing 258 cells, each representing a specific sentencing range. Judges retained some discretion within these ranges, but any departure had to be explicitly justified. Many judges resisted the *Guidelines* as an infringement on their professional autonomy, and many lower court rulings challenged their constitutionality. However, the Supreme Court ultimately upheld them. As Schauer notes, the *Guidelines* were introduced because there was little reason to believe that judges were particularly aware of the frequency or severity of their errors in assessing individual cases (Schauer, 2003, p. 260). The corollary for the present discussion is that if such rudimentary, paper-based tools were considered an improvement over purely human judgment, then today's sentencing algorithms – enhanced by machine learning and trained on large datasets – undeniably represent an even more beneficial improvement.[9] In other words, if courts must make decisions about individuals like Mr. Blue, there is little reason not to employ the best available actuarial tools to support that process.

Equal weight. When it comes to Singer's requirement of assigning equal weight to similar interests, critics of sentencing algorithms would likely argue, as we have seen, that the offender's interest in a fully individualized evaluation is given less weight than society's interest in safety and the efficiency of its judicial system. This claim should be rejected for at least three reasons.

First, the state has a legitimate right to assess the risk of reoffending before granting parole. Whether this assessment is made by a judge, a parole board or an algorithm, the aim remains the same: to estimate the likelihood of future societal harm. The data used in such assessments are essentially the same regardless of who (or what) conducts the analysis. Sentencing algorithms, in fact, may lead to a lesser violation of the equal weight requirement because they can often generate a more detailed and comprehensive picture of the offender than a judge might. Even critics acknowledge that "the individualist objection seems less powerful against ML [machine learning] tools […] than it is against traditional recidivism prediction tools", primarily because they "can be given a very large set of data, and so can make finer-grained predictions" (Davies & Douglas, 2022, p. 103). The use of sentencing algorithms, therefore, does not necessarily entail a greater disregard for the individuality of the offender or their interests than human judgment does. The offender's interests in liberty and individualized treatment are simply balanced against the equally significant interests of the public, such as their interest in an effective judicial system and, especially, in not becoming victims of violence, theft, or other serious harm in the event of a mistaken release. As long as properly designed algorithms can support this balancing process more efficiently, consistently, and impartially than human judges can, Singer's equal weight requirement appears to pose no obstacle to their use. Moreover, it is worth noting a certain structural symmetry in this context: algorithms do not merely determine the fate of the offender, but, through the very success or failure of their

Bagaric and Hunter (2022, pp. 132–133) thus note that humans outperform computers in assessing witness credibility, which depends on interpreting demeanor and body language. In recidivism risk assessment, however, computers are more reliable, as they analyze binary, data-driven factors.

[9] This raises the standard question: At what point does an AI system perform well enough to justify its broader adoption, despite the risks involved? This issue need not concern us here. However, in practice, thresholds for public and institutional acceptance tend to crystallize over time, as seen in the growing reliance on autonomous systems in high-stakes fields such as aviation, transportation and medicine. There is no compelling reason to believe that the judicial context will be an exception, especially since, as shown by Zerilli (2022), judges appear to be less prone to uncritically following algorithmic recommendations.

predictions, also shape the condition of the society into which the offender may (or may not) be released.

Second, particular circumstances shift the relative weight of the interests involved. Let us illustrate this with Singer's example of two earthquake victims and two shots of morphine (Singer, 2011, p. 22). One victim has a severely injured leg and is in agony; the other has a slightly injured thigh and is in mild pain. In this case, Singer argues, equal consideration of interests does not require equal treatment (one shot of morphine for each) but a proportionate response to the weight of their actual interests (two shots for the more severely injured victim). Similarly, in criminal justice, the weight of an offender's interest in early release or, for that matter, an individualized assessment is not absolute. Once a person has been fairly convicted – especially of a serious crime – the weight of their interests may plausibly be reduced, particularly when set against the public interest in preventing further offenses and harm. In such cases, assigning lesser weight to some of the offender's interests (including the one in fully individualized sentencing) does not violate the equal weight requirement. It merely reflects the broader ethical context in which competing interests must be weighed.

Third, there is a difference between contexts in which individualized assessment is indispensable and contexts in which it is less important. Individualized assessment and sentencing are essential when determining guilt for a specific crime. In such cases, the judicial process must examine the unique characteristics of the case and the person accused, ensuring that responsibility is assigned correctly. However, it is far from clear that the same level of individualization is equally important when assessing the risk of recidivism. The decision-making context changes once guilt has already been established through a fair trial. What is now at stake is not the determination of *past* guilt but the prediction of *future* behavior. Unlike guilt, which is tied to concrete past actions and facts, the likelihood of reoffending is always a probabilistic estimation – one that, as we have seen, is often better informed by broader patterns and comparative data than by purely individual traits. This shift in context also justifies a shift in evaluative methods: while individualized scrutiny is crucial for attributing guilt, it may not be equally crucial for forecasting the future. In such cases, using an algorithm to support judges' decision-making (remember that it is always about algorithms supporting judges, not passing judgments on their own) does not mean that the offender's interests are unjustly neglected or given less moral weight. It is only that different questions call for different forms of decision-making. Figuratively speaking, when Mr. Blue stands accused of an armed robbery, his guilt must be established through an individualised process. However, when he applies for parole after serving part of his sentence, statistical considerations are not only permissible but, in fact, unavoidable.[10]

To illustrate the above points, imagine again that Mr. Blue is applying for a job as a military pilot and that two distinct types of medical evaluations are required. One type of evaluation concerns his current health status and aims to determine whether he is fit for the job. This requires an individualized, case-specific assessment involving a thorough examination of his unique physical condition. The other type of evaluation estimates his long-term health prospects – such as the likelihood of developing a particular illness over the next ten years. This kind of assessment is typically based on probabilistic reasoning, using models that compare Mr. Blue's characteristics (e.g., family history and lifestyle) with large datasets of similar individuals. Such

[10] An interesting real-world example of the courts' adherence to an individualised approach in determining guilt is the case of identical twins Hassan and Abbas O., who were suspected of committing a jewelry heist in Berlin in 2009. DNA evidence found in a glove at the crime scene matched both brothers. However, since they shared virtually identical DNA, investigators were unable to determine which one had actually been present. Although one (or both) of them was almost certainly involved, the inability to individualise guilt led to their acquittal (cf. Himmelreich, 2009). Had they been convicted and later applied for parole, however, such strict individualisation would no longer have been required: their risk of reoffending could justifiably have been assessed on statistical grounds alone.

an approach is generally accepted as appropriate, especially in contexts involving public safety or long-term institutional investment. Suppose this assessment reveals a high probability of developing a condition that could, at some point, compromise job performance or public safety, and Mr. Blue is not selected. No moral wrong was committed, provided the process was fair and the data used were relevant. The same applies to sentencing algorithms.

Impartiality. The pessimistic conclusion that sentencing algorithms are incompatible with the impartiality requirement from Singer's principle of equal consideration of interests is also unwarranted. To understand why, we must recall the growing incentive across judicial systems worldwide to introduce such tools. This trend is driven not only by constant technical advancements – such as increasing predictive accuracy, improved calibration, and efforts to mitigate issues like non-transparency and bias – but also by the widely shared view that the *status quo*, in which decisions are made solely by judges or parole boards, is a less desirable alternative. Judges and parole boards, being human, are often influenced by their uniquely human biases – shaped by culturally embedded assumptions, personal upbringing, emotional fluctuations, or even something as trivial as whether they have had lunch.[11] Sentencing algorithms, by contrast, offer the prospect of greater consistency and objectivity in judicial assessments and, by extension, a higher degree of impartiality. As Bagaric and Hunter note, the crucial difference between the (human) *status quo* and the (algorithmic) *status novus* lies in the fact that "it is possible to run regressions over the data used in these algorithms" and thereby "weed out bias", while "there is little that can be done to negate subconscious bias" in judges (Bagaric & Hunter, 2022, p. 137).

A way to uphold the alignment between sentencing algorithms and impartiality requirement is to draw on the well-known distinction between critical and intuitive moral reasoning, originally proposed by R. M. Hare (1981) and later adopted by Singer (2011). On this view, moral thinking operates on two levels: the critical level, marked by slow, reflective and consequence-sensitive deliberation, and the intuitive level, which, due to limited time, information and cognitive capacities, governs most everyday decisions through general and simple moral rules. As both Hare and Singer emphasize, in ordinary circumstances, it is best to rely on intuitive reasoning precisely because critical-level deliberation in such settings is prone to fail due to emotional factors, cognitive bias, self-interest, or time pressure. Since critical reasoning is meant to control and, if necessary, revise our intuitions (not the other way around), the lesson for judicial practice is clear: since judges and parole boards often make decisions under intuitive-level influences, algorithmic decision-making – an alternative that is more systematic and closer in spirit to critical reasoning – offers a more secure path toward enhancing impartiality in sentencing.

Finally, one of the most evident positive impacts of sentencing algorithms is their potential to preserve consistency in decision-making – an aspect closely tied to the ideal of impartiality. Justice, as its traditional statuesque representation suggests, must be blind to irrelevant individual differences and focused solely on the relevant aspects of each case. This ideal is often captured by the (Aristotelian) maxim that "equal cases should be treated equally and unequal cases unequally". Achieving such impartiality requires diachronic and synchronic consistency in sentencing. Diachronic consistency means that judges must be aware of how similar cases have been decided in the past, including the reasoning behind those decisions and their eventual

[11] This phenomenon – namely, that judges tend to make more lenient decisions after a meal break – has become known as the "hungry judge effect", originally proposed in the study by Danziger, Levav and Avnaim-Pesso (2011). Although this particular study has been methodologically challenged from several angles (Weinshall-Margel & Shapard, 2011; Glöckner, 2016; Chatziathanasiou, 2022), the broader point surely remains intact: human judges are vulnerable to various forms of bias that are difficult to detect or correct. Algorithms are immune to such human contingencies, and if, as Singer suggests (2011, p. 11), ethics requires impartial judgment from a universal standpoint, they could represent a more consistent form of ethical reasoning (though rooted *in silico* rather than *in vivo*).

outcomes. Sentencing algorithms support this by being trained on extensive databases of past cases, thus helping to maintain coherence in legal reasoning over time. It is impossible to treat cases as similar – or even to recognize them as such – without a reliable understanding, grounded in numerous past examples, of what makes them similar or dissimilar in the first place. These disparities have long been a source of concern and have prompted various institutional responses – such as the introduction of the already mentioned *Federal Sentencing Guidelines* – aimed at reducing judicial discretion and promoting more uniform outcomes. By applying consistent criteria across all cases, sentencing algorithms are expected to reduce such disparities further and thereby contribute to an impartial justice system.[12]

No irrelevant groups. Singer is right when he claims that, when it comes to the principle of equal consideration of interests, group classifications such as race, sex, or IQ are as "irrelevant to the undesirability of pain" as any trivial group classification, such as being born in a leap year or having more than one vowel in one's surname (Singer, 2011, p. 22). The principle, as he emphasizes, "is strong enough to rule out an intelligence-based slave society as well as cruder forms of racism and sexism" (Singer, 2011, p. 22). Does the same "no irrelevant groups" logic apply to sentencing algorithms and their classifications? Can we reject them in the same way as Singer rejects classifications based on race, sex, IQ, being born in a leap year, or having more than one vowel in one's surname? We cannot.

It should be emphasized that some attributes, such as race and ethnicity, are widely regarded as inadmissible in sentencing algorithms used in courts. Moreover, even attributes that could potentially serve as proxies for such characteristics (e.g., postal code or gang affiliation) are typically subjected to stricter auditing procedures. What is more important in this context, however, is the following: Singer's "no irrelevant groups" requirement is certainly sound when the basic utilitarian currency – the undesirability of pain and the desirability of pleasure (broadly construed) – is at stake. It is undeniable that pain is bad regardless of the group to which the individual experiencing it belongs, and that, if interests ought to be weighed equally, one individual's pain matters as much as another's. However, this line of reasoning is not applicable to the issue of sentencing algorithms, because their central aim is to estimate the probability that an individual will reoffend in the future. To return to our earlier example, what matters most is not whether a given decision satisfies or impedes Mr. Blue's preferences, but whether society will be at greater or lesser risk if he is released. In recidivism predictions, it makes a difference to know, for example, whether an individual is a repeat offender with a history of substance abuse, domestic violence, and impaired behavioral control or a first-time offender with a stable family background and no substance abuse record. For such purposes, certain group-based classifications and comparisons are relevant, just as they are in actuarial assessments for life insurance.

As already noted, the criteria used by sentencing algorithms rely on empirically validated regularities rather than arbitrary stereotypes. This aligns closely with Singer's naturalistic conception of human nature, as articulated across different phases of his thought (Singer, 1981; 1999), where he defends a view of human behavior consistent with the findings of evolutionary psychology. He rejects the radical (leftist) claim that human nature is either nonexistent or so malleable that it can be entirely reshaped through social reforms. Instead, he argues that "we are evolved animals, and that we bear the evidence of our inheritance, not only in our anatomy and our DNA, but in our behavior too" (Singer, 1999, p. 6). This evidence includes behavioral and psychological traits such as loyalty to kin, various forms of cooperation, hierarchical structures, sex roles, ethnic identification, xenophobia or racism. Singer cautions that we cannot "expect to end all conflict and strife between human beings, whether by political revolution,

[12] Lippert-Rasmussen (2011), although skeptical about sentencing algorithms (cf. Lippert-Rasmussen, 2022), allows that statistical discrimination need not always be inherently wrong and that its possible wrongness need not stem from a failure to treat people as individuals.

social change, or better education", insisting that "policies can be grounded on the best available evidence of what human beings are like" (Singer, 1999, p. 61). This is highly relevant when it comes to possible justification of sentencing algorithms, which can be understood as technological extension of precisely such empirically informed policies, i.e., as tools built on the best available data about human behavior.[13]

No system of parole decision-making is perfect. Since these systems aim to predict future behavior, they are inevitably marked by a high degree of uncertainty. However, uncertainty is only part of the problem. Equally important are the fairness-related trade-offs that arise from the limitations of any predictive arrangement. In the case of judges or parole boards, such trade-offs stem from distinctively human constraints: limited cognitive resources, time pressure, and the influence of conscious or unconscious biases. Sentencing algorithms are not without their own difficulties. They come with their own set of trade-offs, including the risk of perpetuating biases present in historical data they are trained on, or misclassifying certain applicant groups due to suboptimal predictive accuracy or calibration. For this reason, under any defensible conception of fairness (including one informed by Singer's principle of equal consideration of interests), it is reasonable to expect that rigorous technical and procedural safeguards accompany such tools. These should probably include external validation, performance monitoring and subgroup calibration checks, documented expert oversight, and, crucially, an accessible right to appeal recommendations. In short, the claim is not that such protections become unnecessary when algorithms are used. Rather, the point is that in systems incorporating sentencing algorithms, such safeguards may be more easily implementable and more reliably enforced than in systems based exclusively on unaided human judgment.

Conclusion

Sentencing algorithms are not only consistent with Peter Singer's preference utilitarianism and the principle of equal consideration of interests but, in many respects, supported by them. Although Singer did not systematically address sentencing algorithms, such a conclusion is relevant and illuminating. The lasting influence of Singer's ethical position stems, arguably, from its combination of two key features: a commitment to making the world a better place in a way that is grounded in reason and practical feasibility, and a concern for individual interests and the avoidance of unjust discrimination. These values resonate not only with utilitarians but also with ethicists holding differing views and, perhaps more importantly, with the broader public. If, as this paper hopefully demonstrates, Singer's ethical framework can accommodate sentencing algorithms, it could be interpreted as a meaningful – moreover, reciprocal – endorsement for both sentencing algorithms and Singer's preference utilitarianism.

Acknowledgments

This paper was produced with the support of the Croatian Science Foundation (HRZZ) as part of the *Artificial Intelligence, Autonomy and Justice* project conducted at the Institute of Philosophy in Zagreb (project number: IP-2022-10-1130). I am grateful to Neven Sesardić and to two anonymous reviewers for reading the manuscript and providing valuable comments.

The Sekyra Foundation funds the open access publication of the article in the journal.

[13] That Singer would be open to predictions based on relevant empirical evidence is indirectly suggested by a claim he makes in the context of animal protection: "Many studies in psychology and criminology have shown that violent criminals are likely to have a history of animal abuse during their childhood or adolescence" (Singer, 2011, p. 67). Given this, it seems likely that if Singer had a say in the design of sentencing algorithms, he might propose including animal abuse as a factor in assessing recidivism risk.

References

ANGWIN, J., LARSON, J., MATTU, S. & KIRCHNER, L. (2016): Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. In: *ProPublica*, [online] [Retrieved August 25, 2025] Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

BAGARIC, M. & HUNTER, D. (2022): Enhancing the integrity of the sentencing process through the use of artificial intelligence. In: J. Ryberg & J. V. Roberts (eds.): *Sentencing and Artificial Intelligence*. New York: Oxford University Press, pp. 122–144.

CHATZIATHANASIOU, K. (2022): Beware the lure of narratives: "Hungry judges" should not motivate the use of "artificial intelligence" in law. In: *German Law Journal*, 23(4), pp. 452–464.

CHIAO, V. (2024): Algorithmic decision-making, statistical evidence and the rule of law. In: *Episteme*, 21(4), pp. 1241–1265.

DANZIGER, S., LEVAV, J. & AVNAIM-PESSO, L. (2011): Extraneous factors in judicial decisions. In: *Proceedings of the National Academy of Sciences of the United States of America*, 108(17), pp. 6889–6892.

DAVIES, B. & DOUGLAS, T. (2022): Learning to discriminate: The perfect proxy problem in artificially intelligent sentencing. In: J. Ryberg & J. V. Roberts (eds.): *Sentencing and Artificial Intelligence*. New York: Oxford University Press, pp. 97–121.

FLORES, A. W., LOWENKAMP, C. T. & BECHTEL, K. (2017): False positives, false negatives, and false analyses: A rejoinder to 'Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks". In: *Federal Sentencing Reporter*, 30(1), pp. 27–32.

GHOSE, S., TSE, Y. F., RASAEE, K., SEBO, J. & SINGER, P. (2024): The case for animal-friendly AI. arXiv preprint arXiv:2403.01199

GHOSE, S., HÄYRY, M. & SINGER, P. (2025): Sentience and beyond – A representative interview with Peter Singer AI. In: *Cambridge Quarterly of Healthcare Ethics*, First View, pp. 1–9.

GLÖCKNER, A. (2016): The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. In: *Judgment and Decision Making,* 11(6), pp. 601–610.

HAGENDORFF, T., BOSSERT, L. N., TSE, Y. F. & SINGER, P. (2023): Speciesist bias in AI: how AI applications perpetuate discrimination and unfair outcomes against animals. In: *AI and Ethics* 3, pp. 717–734.

HARE, R. M. (1981): *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.

HIMMELREICH, C. (2009): Despite DNA evidence, twins charged in heist go free. In: *Time*, 23 March, [online] [Retrieved August 25, 2025] Available at: https://time.com/archive/6946089/despite-dna-evidence-twins-charged-in-heist-go-free

LIPPERT-RASMUSSEN, K. (2011): "We are all different": Statistical discrimination and the right to be treated as an individual. In: *The Journal of Ethics*, 15(1), pp. 47–59.

LIPPERT-RASMUSSEN, K. (2022): Algorithm-based sentencing and discrimination. In: J. Ryberg & J. V. Roberts (eds.): *Sentencing and Artificial Intelligence*. New York: Oxford University Press, pp. 74–96.

MARR, B. & WARD, M. (2019): *Artificial Intelligence in Practice: How 50 Successful Companies Used AI and Machine Learning to Solve Problems*. Hoboken, NJ: Wiley.

REICH, C. L. & VIJAYKUMAR, S. (2021): A possibility in algorithmic fairness: Can calibration and equal error rates be reconciled? In: *2nd Symposium on Foundations of Responsible Computing (FORC 2021). Leibniz International Proceedings in Informatics (LIPIcs)* 192, pp. 4: 4:1–4:21.

SCHAUER, F. (2003): *Profiles, Probabilities and Stereotypes*. Cambridge, MA. & London: Belknap Press.
SCHWARZE, M. & ROBERTS, J. V. (2022): Reconciling artificial and human intelligence: Supplementing not supplanting the sentencing judge. In: J. Ryberg & J. V. Roberts (eds.): *Sentencing and Artificial Intelligence*. New York: Oxford University Press, pp. 206–225.
SIEGEL, E. (2018): *Predictive Analytics: The Power to Predict Who Will Click, Lie, or Die*. Hoboken, NJ: Wiley.
SINGER. P. (1981): *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton & Oxford: Princeton University Press.
SINGER, P. (1999): *A Darwinian Left: Politics, Evolution, and Cooperation*. London: Weidenfeld & Nicolson.
SINGER, P. (2011): *Practical Ethics*. New York: Cambridge University Press.
SINGER, P. (2023): *Ethics in the Real World: 90 Essays on Things That Matter*. Princeton & Oxford: Princeton University Press.
SINGER, P. & TSE, Y. F. (2023): AI ethics: the case for including animals. In: *AI and Ethics*, 3, pp. 539–551.
WEINSHALL-MARGEL, K. & SHAPARD, J. (2011): Overlooked factors in the analysis of parole decisions. In: *Proceedings of the National Academy of Sciences*, 108(42), p. E833.
ZERILLI, J., DANAHER, J., MACLAURIN, J., GAVAGHAN, C., KNOTT, A., LIDDICOAT, J. & NOORMAN, M. (2021): *A Citizen's Guide to Artificial Intelligence*. Cambridge, MA: The MIT Press.
ZERILLI, J. (2022): Algorithmic sentencing: Drawing lessons from human factors research. In: J. Ryberg & J. V. Roberts (eds.): *Sentencing and Artificial Intelligence*. New York: Oxford University Press, pp. 165–183.